COVID-19 DATA ANALYSIS AND FORECASTING

Name: Reahaan Sheriff I

Reg no: 2019202045

Degree: MCA 3yrs Reg

Project Guide: Ms.P.S Apirajitha

Abstract

COVID-19 has sparked a worldwide pandemic, with the number of infected cases and deaths rising on a regular basis. Along with recent advances in soft computing technology, researchers are now actively developing and enhancing different mathematical and machine-learning algorithms to forecast the future trend of this pandemic. Thus, if we can accurately forecast the trend of cases globally, the spread of the pandemic can be controlled. In this project, a LSTM model will be used on a time-series dataset to forecast the cases of COVID-19 in future.

Introduction

Machine learning (ML) is a category of an algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available.

Long Short Term Memory is a kind of recurrent neural network. In RNN output from the last step is fed as input in the current step. It tackled the problem of long-term dependencies of RNN in which the RNN cannot predict the word stored in the long-term memory but can give more accurate predictions from the recent information. As the gap length increases RNN does not give an efficient performance. LSTM can by default retain the information for a long period of time. It is used for processing, predicting, and classifying on the basis of time-series data.

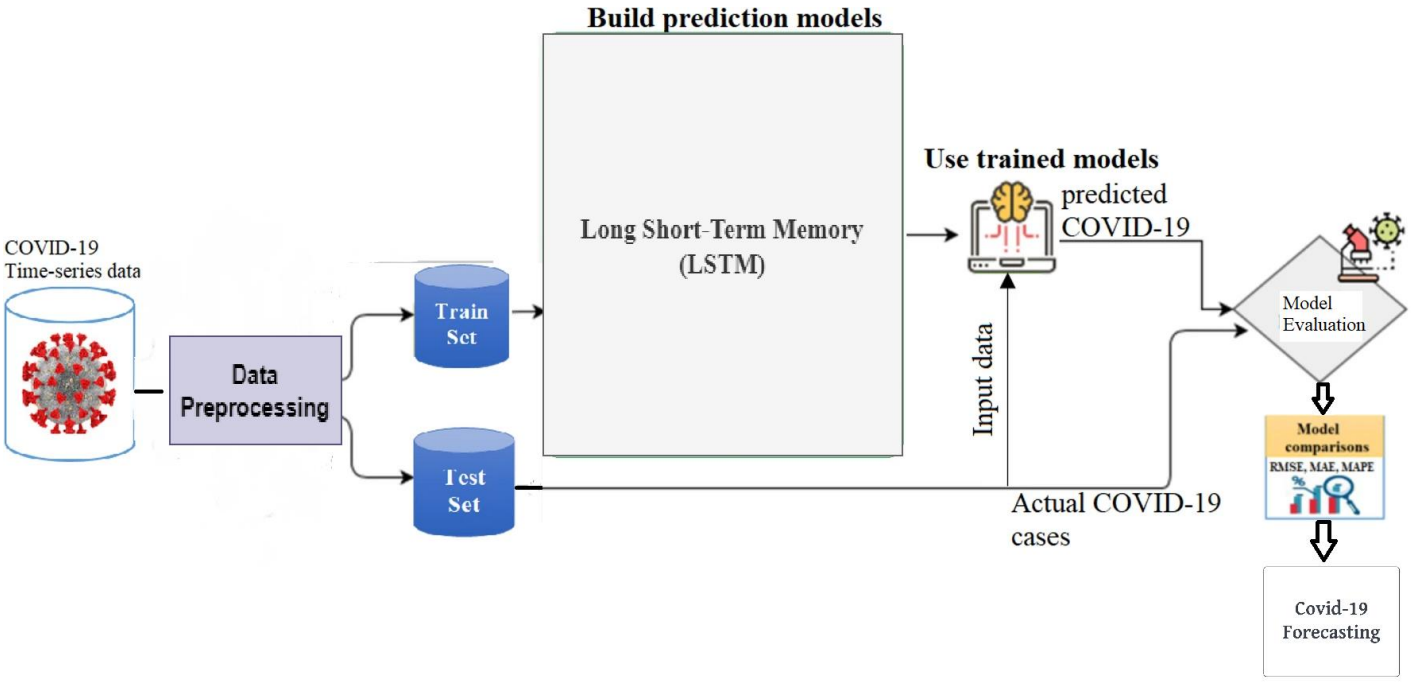
Problem statement

The COVID-19 disease has changed the global landscape completely. A high reproduction rate and a higher chance of complications have led to border closures, empty streets, rampant stockpiling, mass self-isolation policies, and an economic recession. We will go through the process of performing Exploratory Data Analysis (EDA) on COVID-19 global data to forecast active cases, cases of recovery, and death. We have used Long Short-Term Memory (LSTM) architecture, a Deep Learning technique for building the model.

Objective

To predict the cases and peak case date for COVID-19 in India as accurately as possible. We will be using LSTM networks.

Architecture diagram



Architecture diagram explanation

1. Raw data - Collect real time data set for covid 19 cases
2. Pre- processing – Cleaning the data and split data in training and testing sets,
3. Time series modeling – construct machine learning LSTM to model the covid 19 data
4. Forecasting – Forecast future of covid 19 using the constructed models, Evaluate forecasting accuracy using statistical indicators.
5. Visualize the results.

List of modules

- Data pre-processing
- Building model
- Training model
- Model evaluation
- Visualization and forecasting

Brief description about modules

Data pre-processing

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

Building model

We will built the model with the help of LSTM. The model has an input layer followed by three LSTM layers. The LSTM layers contain Dropout as 0.5 to prevent overfitting in the model. The output layer consists of a dense layer with 1 neuron with activation as relu. We will predict the number of Corona cases, so our output will be a positive number $(0, \infty)$.

Training model

To train the model, we will take out training data (80%) and used 20% of it as validation data. To lower the learning rate of our model we will use reducelronplateau in the model. Training the model with n epochs.

Model evaluation

It estimates how well (or how bad) the model is, in terms of its ability in mapping the relationship between X (a feature, or independent variable, or predictor variable) and Y (the target, or dependent variable, or response variable).

Visualization and forecasting

In order to see the prediction and accuracy, first, we predicted the output of our x_test data. This was the output that we got from the test data. To accurately plot the values, we needed to bring our prediction and y_test data back to the original bounds of the data. In the end, we plotted a graph between the actual COVID-19 cases compared to our predicted COVID-19 cases to see the overall accuracy of our model.

References

1. Wang J., Liu Y., Wei Y., Xia J., Yu T., Zhang X., Zhang L. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet. 2020;395(10223):507–513.

2. JA Backer, D Klinkenberg and J. Wallinga, Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020.

3. Tolga Ergen and Suleyman Serdar Kozat, "Efficient online learning algorithms based on LSTM neural networks", IEEE transactions on neural networks and learning systems, vol. 29, no. 8, pp. 3772-3783, 2017.

4. H. Li, S.-M. Liu, X.-H. Yu, S.-L. Tang and C.-K. Tang, "Coronavirus disease 2019 (COVID-19): current status and future perspectives", International Journal of Antimicrobial Agents, pp. 105951, 2020.

Project guide